System, Method and Computer Program Product For Fast and Efficient Searching of Large Chemical Libraries

Inventors: Victor S. Lobanov
Dimitris K. Agrafiotis
Francis R. Salemme

Background of the Invention

Field of the Invention

The present invention relates generally to searching of virtual combinatorial libraries. More particularly, the present invention relates to the selection of compounds, based on fitness functions, from large virtual combinatorial libraries.

Related Art

The explosive growth of combinatorial chemistry in recent years has been greeted as both a blessing and a curse. While it has solved the problem of throughput and has allowed the traditionally slow drug discovery process to be conducted in a massively parallelized fashion, it has created the need to deal with compound collections of truly staggering size. These include both physical collections of compounds that are synthesized using automated parallel synthesis, as well as virtual collections containing molecules that could potentially be synthesized by systematic application of established synthetic principles. The initial ambition to 'make and test them all' has given way to a more pragmatic approach once it became evident that 'all' was a number of immense proportions. For example, a simple diamine-based combinatorial library built from only commercially available reagents can include up to 10^{12} compounds which is equivalent to approximately 300 years of synthesis and testing at a rate of 10 million compounds per day (Cramer *et al.*, "Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research," J. Chem. Inf.

Comput. Sci. 1998, 38, 1010-1023). The recognition of these practical limitations and the desire to use the available synthetic and screening resources in an efficient way has generated interest in virtual chemistry and, in particular, systems and methods for handling and analyzing large chemical libraries. More specifically, for example, there is an interest in efficiently selecting compounds that are similar to a particular query structure (e.g., drug lead) or selecting compounds that have desired properties.

Searching a virtual combinatorial library for compounds that are similar to a particular query structure (or query structures) or have a set of desired properties typically involves three steps for each compound: enumeration, calculation of descriptors, and evaluation of similarity or estimation of the property of interest. Due to the large number of possible products in many virtual combinatorial libraries (particularly three- and four-component ones), just the enumeration part alone can take a few weeks of computational time. Additionally, the storage requirements for a fully enumerated virtual combinatorial library can be prohibitive. Since in these cases neither the generation nor the storage of fully enumerated libraries and their associated descriptors is feasible, there is a need for systems and methods that can identify the desired compounds without enumerating the entire library.

One possible solution is to look at the far less numerous reagents instead of the products. The reagent-based approach is frequently used to maximize molecular diversity, and is based on the assumption that diverse reagents will lead to diverse products. However, it was recently shown that a selection based on the products themselves can be substantially more diverse, perhaps by as much as 35-50% (Gillet *et al.*, "The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries," J. Chem. Inf. Comput. Sci. 1997, 37, 731-740). When the selection criterion is similarity, the final products themselves must be considered, and the only proposed solution has been to use additive or otherwise "decomposable" descriptors. These are descriptors which, for combinatorial products, can be computed from the values of the corresponding

descriptors of their constituent reagents (Cramer *et al.*, "Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research," J. Chem. Inf. Comput. Sci. 1998, 38, 1010-1023).

Thus, the need remains for a system and method for efficiently and effectively generating product-based selections from large virtual combinatorial libraries. More generally, there is a need for a system and method for efficiently and effectively searching large virtual combinatorial libraries based on a fitness function.

There do exist some virtual combinatorial libraries that have already been fully or partially enumerated. However, there is currently a deficiency of satisfactory systems and methods for efficiently and effectively searching these enumerated virtual combinatorial libraries based on a fitness function. Accordingly, there is also a need for a system and method for efficiently and effectively searching large enumerated virtual combinatorial libraries based on a fitness function.

Summary of the Invention

The present invention is a system, method, and computer program product for searching of large virtual combinatorial libraries based on a fitness function. Conventional systems and methods for searching virtual combinatorial libraries typically enumerate and characterize every reagent combination (i.e., possible compound) associated with the virtual combinatorial library, characterize every enumerated compound, and then perform an evaluation, based on a fitness function, for every compound. Selecting compounds from a large virtual combinatorial library using such conventional systems and methods requires prohibitively excessive amounts of time and resources. The present invention reduces the amount of time and resources that are necessary to search a large virtual combinatorial library by not requiring the enumeration, characterization, and evaluation of every reagent combination associated with the library.

According to the method of an embodiment of the present invention, a first set of N reagent combinations are selected from a virtual combinatorial library. The selection can be random or non-random. Each reagent combination in the first set is then enumerated to produce a first set of enumerated compounds. M number of compounds of the first set of enumerated compounds are selected based on a fitness function. The M compounds are then deconvoluted into reagents to generate a focused library. Every reagent combination associated with the focused library is enumerated to produce a second set of enumerated compounds. K number of compounds of the second set of enumerated compounds are then selected based on the fitness function. These K compounds represent a near optimal selection of compounds based on the fitness function.

Examples of fitness functions that can be used with the present invention include, but are not limited to, similarity to one or more query structures, diversity, and presence of desired properties. That is, for example, the present invention can be used to efficiently and effectively select, from a large virtual combinatorial library, a set of the near optimal most similar compounds to a drug lead. The present invention can also be used to efficiently and effectively select a diverse collection of compounds from a large virtual combinatorial library. Additionally, the present invention can be used to select compounds having desired properties from a large virtual combinatorial library.

In an alternative embodiment, the present invention can be used to search large enumerated virtual combinatorial libraries. This embodiment of the present invention takes advantage of those virtual combinatorial libraries that have already been enumerated.

Features and advantages of the present invention, as well as the structure and operation of various embodiment of the present invention, are described in detail below with reference to the accompanying drawings. In the drawings, like reference numbers indicate identical or functionally similar elements. Also, the leftmost digit(s) of the reference numbers identify the drawings in which the associated elements are first introduced.

Brief Description of the Figures

The present invention will be described with reference to the accompanying drawings, wherein:

FIG. 1 is a process flowchart illustrating a method for searching of virtual combinatorial libraries, according to an embodiment of the present invention;

FIG. 1A is a process flowchart illustrating a method for searching of enumerated virtual combinatorial libraries, according to an embodiment of the present invention;

FIGS. 2 and 3 show a matrix that represents a small portion of a large virtual combinatorial library;

FIG. 4 shows a matrix that represents a small portion of a focused library that is generated using the present invention;

FIG. 5 is a process flowchart illustrating a method for similarity searching of virtual combinatorial libraries, according to an embodiment of the present invention;

FIG. 6 is a process flowchart illustrating a method for similarity searching of virtual combinatorial libraries, wherein the total number of reagents that make up the final collection of similar compounds are reduced, according to an embodiment of the present invention;

FIG. 7 is a process flowchart illustrating a method for dissimilarity (diversity) searching of virtual combinatorial libraries, according to an embodiment of the present invention;

FIG. 8 is a process flowchart illustrating a method for dissimilarity (diversity) searching of virtual combinatorial libraries, wherein the total number of reagents that make up the final collection of dissimilar compounds is reduced, according to an embodiment of the present invention;

FIG. 9a illustrates a synthetic protocol for a diamine virtual combinatorial library that was used to demonstrate the effectiveness of the present invention;

FIG. 9b illustrates a synthetic protocol for a Ugi reaction that was also used to demonstrate the effectiveness of the present invention;

FIG. 10a shows a query structure that was used to demonstrate the effectiveness of the present invention;

FIG. 10b shows another query structure that was used to demonstrate the effectiveness of the present invention;

FIG. 11a is a graph that illustrates a similarity profile of the diamine virtual combinatorial library of FIG. 9a;

FIG. 11b is a graph that illustrates a similarity profile of the Ugi virtual combinatorial library of FIG. 9b;

FIGS. 12a and 12b are graphs that illustrate the overlap between stochastic selections, generated using the present invention, and reference selections;

FIG. 13 is a table that summarizes the experimental results obtained when using an embodiment of the present invention;

FIGS. 14a, 14b, 15, 16a, 16b and 17 are graphs that show how the selection of particular variables affects experimental results of the present invention;

FIG. 18 shows the structures of some of the most similar compounds found during experimental testing of the present invention;

FIG. 19 shows an exemplary environment in which the present invention can be used;

FIG. 20 shows an example of a computer system that can be used to implement the present invention;

FIG. 21 shows examples of diamines, and acidchlorides and halocarbons (i.e., alkylating/acylating agents), associated with the diamine virtual combinatorial library that was used to demonstrate the effectiveness of the present invention; and

FIG. 22 shows examples of acids, amines, aldehydes, and isonitriles, associated with the Ugi virtual combinatorial library that was used to demonstrate the effectiveness of the present invention.

Detailed Description of the Preferred Embodiments

Table of Contents

1. General Overview

The present invention is directed to a system, method, and computer program product for searching large virtual combinatorial libraries based on fitness functions. In one embodiment, the fitness function is similarity to one or more query structures (i.e., probe(s)). In this embodiment, the present invention can be used to significantly reduce the amount of time and resources that it takes to search a large virtual combinatorial library for a set of compounds that are similar to a query structure(s) (e.g., a drug lead). In another embodiment, the fitness function is related to diversity of a collection of compounds. In the diversity embodiment, the present invention can be used to significantly reduce the amount of time and resources that it takes to search a large virtual combinatorial library for a collection of diverse compounds.

The fitness function can also be related to a desired characteristic. That is, the present invention can be used to significantly reduce the amount of time and resources that it takes to search a large virtual combinatorial library for compounds that exhibit (or do not exhibit) specific characteristics. Such characteristics can include, for example, physical properties, chemical properties, functional properties and/or bioactive properties, although the invention is not limited to these characteristics.

According to one specific embodiment, the present invention is directed to a system, method, and computer program product for generating product-based similarity selections from large virtual combinatorial libraries. An advantage of this embodiment of the present invention, as compared to conventional similarity searching techniques, is that the present invention does not require enumeration and descriptor generation for every reagent combination associated with a virtual combinatorial library. This results in a significant reduction in the resources necessary to carry out similarity selections. Additional advantages of this embodiment of the present invention is that it is not limited to additive or "decomposable" descriptors. Further, this embodiment of the present invention

is able to provide an optimal or nearly optimal similarity selection in a reasonable time frame. For illustrative purposes only, this embodiment will be described in the most detail. However, the intention is not to limit the present invention to use in similarity searching. Rather, the intention is to provide sufficient detail so that one skilled in the art can use the present invention with various fitness functions. Alternative fitness functions include, but are not limited to, diversity and presence of one or more desired properties.

Conventional systems used for similarity selection enumerate and characterize each reagent combination (i.e., possible compound) associated with a virtual combinatorial library, characterize every enumerated compound using descriptors, and then perform a similarity evaluation for every possible compound. Performing similarity selection using such conventional systems is extremely time inefficient. For example, the enumeration, characterization, and similarity evaluation of a virtual combinatorial library containing 6.75 million possible compounds (i.e., reagent combinations) required 34 hours on a dual processor 400 MHz Intel Pentium II machine.

In contrast, using the present invention, a similarity evaluation of the same 65 million possible compounds, using the same dual processor 400 MHZ Intel Pentium II machine, required only 30 minutes. This large reduction in time is due to the fact that the present invention does not perform enumeration, characterization, and similarity evaluation for all of the 6.75 million possible compounds.

The present invention is also directed to a system, method, and computer program product for searching large enumerated virtual combinatorial libraries based on fitness functions. This embodiment of the present invention takes advantage of those virtual combinatorial libraries that have already been enumerated.

The concepts of virtual libraries, virtual combinatorial libraries, molecular similarity, enumeration, and the selection problem associated with virtual

combinatorial libraries are of particular pertinence to the present invention. Accordingly, each of these concepts are discussed in some detail below.

Virtual Library. A virtual library is essentially a computer representation of a collection of chemical compounds obtained through actual and/or virtual synthesis, acquisition, or retrieval. By representing chemicals in this manner, one can apply cost-effective computational techniques to identify compounds with desired physico-chemical properties, or compounds that are diverse, or similar to a given query structure. By trimming the number of compounds being considered for physical synthesis and biological evaluation, computational screening can result in significant savings in both time and resources, and is now routinely employed in many pharmaceutical companies for lead discovery and optimization.

Virtual Combinatorial Libraries. Whereas a compound library generally refers to any collection of actual and/or virtual compounds assembled for a particular purpose (for example a chemical inventory or a natural product collection), a virtual combinatorial library represents a collection of compounds derived from the systematic application of a synthetic principle on a prescribed set of building blocks (i.e., reagents). These building blocks are grouped into lists of reagents that react in a similar fashion (e.g. A reagents and B reagents) to produce the final products constituting the library (C, $A_i + B_j \rightarrow C_{ij}$). Full virtual combinatorial libraries encompass the products of every possible combination of the prescribed reagents, whereas sparse combinatorial libraries (also called sparse arrays) include systematic subsets of products derived by combining each A_i with a different subset of B_j's. Unless mentioned otherwise, the term virtual combinatorial library will hereafter imply a full virtual combinatorial library.

A virtual combinatorial library can be thought of as a matrix with reagents along each axis of the matrix. For example, the chemical reaction $A_i + B_j \rightarrow C_{ij}$, may be represented by a two dimensional matrix with the A reagents along one axis and the B reagent along another axis. If there exist 10 different A reagents

and 10 different B reagents, then a virtual combinatorial library representing this chemical reaction would be a 10 x 10 matrix, with 100 possible products (also referred to as possible compounds or reagent combinations). If the chemical reaction to be represented by a virtual library were $A_i + B_j + C_k \rightarrow D_{ijk}$, and reagent class A included 1,000 reagents, reagent class B included 10,000 reagents, and reagent class C included 500 reagents, then a virtual combinatorial library representing this chemical reaction would be a 1,000 x 10,000 x 500 matrix (i.e., a three dimensional matrix), with 5×10^9 possible products (i.e., D_{ijk}s).

The possible products that are represented by cells of the virtual combinatorial library matrix need not be explicitly represented. That is, the possible products in each cell of the matrix need not be enumerated. Rather, the possible products in each cell can simply be thought of as Cartesian coordinates corresponding to a particular reagent combination, such as A_1B_5. Unless mentioned otherwise, a virtual combinatorial library should be thought of as a matrix representing a chemical reaction where the products have not been enumerated. Explained another way, a virtual combinatorial library can be thought of as a matrix having a defined size but with empty cells. Each empty cell can be labeled as a reagent combination (e.g., A_1B_5). In contrast, a fully enumerated virtual combinatorial library can be thought of as a matrix having an enumerated compound in each cell. Unless specifically referred to as an enumerated virtual combinatorial library, mention of a virtual combinatorial library refers to a non-enumerated virtual combinatorial library.

Enumeration. Enumeration is the process of constructing computer representations of a structure of one or more products associated with a virtual combinatorial library. Enumeration is accomplished by starting with reagents and performing chemical transformations, such as making bonds and removing one or more atoms, to construct explicit product structures. In the general sense, enumeration of an entire virtual combinatorial library means explicitly generating

representations of product structures for every possible product of the virtual combinatorial library.

A computer system may require inordinate amounts of time and resources to enumerate every reagent combination associated with a large virtual combinatorial library. Further, the disk requirements for storing every enumerated product (i.e., product structure) can be excessive. Accordingly, a feature of the present invention is to reduce the amount of products that are enumerated during searching of a large virtual combinatorial library.

Virtual Combinatorial Library Generation. Once a synthetic protocol is designed, a virtual combinatorial library can be created. In addition to providing the basis for computational screening, these libraries are also convenient for tracking and archiving purposes. The conceptual approach to generating virtual combinatorial libraries is straightforward. The reaction transformations that convert the reagents into products (i.e., compounds) is reduced into a set of substructure patterns which are mapped onto the reagents to identify reacting groups or atoms, and a list of instructions of how to modify the chemical graphs. These instructions include actions such as removing an existing atom or bond, inserting a new bond between two atoms, or changing the order of a bond. For more complex reactions, the modifications may also include changing the formal charge or chirality of an atom. Complications arise from the fact that the substructure patterns have to be correctly defined so that they map only to those parts of a molecule that would indeed react under the prescribed conditions. For example, if one of the reagents is an amine, it can be defined as a nitrogen atom connected to a carbon and to at least one hydrogen atom (to account for both primary and secondary amines). However, such a definition would also include amides, which are chemically different from amines. Hence, the definition of the amine pattern has to be extended to include further neighboring atoms. Alternatively, substructures which should be avoided can be introduced into the instructions. Furthermore, if primary amines are more reactive than secondary

amines, primary and secondary amines should be mapped separately and assigned different priorities. Additional issues that need to be addressed are removal of protecting or leaving groups (which may or may not be present), and handling of multiple possible products due to regio- or stereo-isomerism or the presence of multiple reactive functionalities within the same reagent.

By their very nature, virtual combinatorial libraries can reach extremely large sizes even with a relatively modest number of building blocks, particularly if there is a large number of variation sites (a problem known as combinatorial explosion). For example, if enumerated, the above-mentioned diamine library can easily include 10^{12} reagent combinations. Were it to be enumerated, this library would contain 50,000 times more compounds than the world's cumulative chemical literature, and would require 10 terabytes of storage space at 100 bytes per structure. Finally, the enumeration of the entire library at a rate of 100,000 structures per second would take over three months. Accordingly, as stated by R.D. Cramer *et al.* in their article, "Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research," J. Chem. Inf. Comput. Sci. 1998, 38, pages 1010-1023, when dealing with virtual combinatorial libraries of this size explicit enumeration is not an option.

Generally, a universally applicable program has been developed for generating virtual libraries. As input, this program takes lists of reagents supplied in Structure-Data File (SDF) format, developed by MDL Information Systems Inc., San Leandro, California, or Simplified Molecular Input Line Entry Specification (SMILES) format, introduced by Daylight Chemical Information System, Inc., Los Altos, California. Reaction definitions can be written in an extension of a scripting language, such as Tool Command Language (Tcl). The use of a scripting language provides a powerful, human-readable, and convenient way of encoding chemical reactions. All chemically feasible transformations are supported, including multiple reactive functionalities, different stoichiometries, cleavage of protecting groups, and many others. The library is stored in a compact format without an explicit enumeration of the products. The

computational requirements of the algorithm are minimal (even a billion-membered library can be generated in a few CPU seconds on a personal computer) and are determined not by the size of the library but by the number of reagents. Despite the implicit encoding, individual structures can be accessed at a rate of 1,000,000 per CPU second.

Use of Virtual Combinatorial Libraries. The role of virtual combinatorial libraries in drug discovery is to provide computational access to compounds that can be readily synthesized and tested for biological activity. The role of the computational tools is to identify which compounds from the library need to be tested to achieve the desired objective. For example, in lead discovery, the objective is to explore different structural classes in order to identify "activity islands" in structure-activity space. Hence, the selected compounds have to be dissimilar from one another so that each compound provides unique and non-redundant information on the structure activity relationship (SAR) landscape. Selecting compounds based on their dissimilarity or diversity has become very popular in recent years, and there has been an extensive number of publications which address this subject. Examples of such publications include: Agrafiotis, "On the Use of Information Theory for Assessing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 576-580; Agrafiotis, "Stochastic Algorithms for Maximizing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 841-851; Agrafiotis, "Diversity of Chemical Libraries," in Encyclopedia of Computational Chemistry, 1998; Clark, "OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets," J. Chem. Inf. Comput. Sci. 1997, 37, 1181-1188; Gillet *et al.*, "The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries," J. Chem. Inf. Comput. Sci. 1997, 37, 731-740; Gillet *et al.*, "Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties," J. Chem. Inf. Comput. Sci. 1999, 39, 169-177; and Martin *et al.*, "Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery," J. Comb. Chem. 1999, 1, 32-45.

Once one or more initial leads have been identified (that is, compounds that show activity against a target and are structurally novel), emphasis is shifted towards exploring more extensively the structure-activity space around those one or more lead molecules. Typically, this is accomplished by selecting and screening compounds that are similar to the initial lead(s). Finally, the accumulated qualitative, and if available quantitative, SAR information is used to optimize the initial leads into preclinical candidates through conventional medicinal chemistry techniques.

Besides similarity and diversity (i.e., dissimilarity), other selection criteria can also be employed. Examples include selecting compounds having desired properties or property distributions as determined by a property prediction algorithm or a quantitative structure-activity model, or exhibiting an optimal fit to a biological receptor as determined by a biomolecular docking algorithm. Compound can also be selected based on 2D and 3D QSAR predictions, and receptor complementarity. Additional details of these and other selection criteria are described in the following patents and patent application, each of which is incorporated herein by reference in its entirety: U.S. Pat. No. 5,463,564, entitled "System and Method of Automatically Generating Chemical Compounds with Desired Properties"; U.S. Pat. No. 5,574,656 entitled "System and Method of Automatically Generating Chemical Compounds with Desired Properties"; U.S. Pat. No. 5,684,711, entitled, "System, Method, and Computer Program for at Least Partially Automatically Generating Chemical Compounds Having Desired Properties"; U.S. Pat. No. 5,901,069, entitled "System, Method, and Computer Program Product for at Least Partially Automatically Generating Chemical Compounds with Desired Properties from a List of Potential Chemical Compounds to Synthesize"; and U.S. Pat. Appl. No.08/963,870, entitled "System, Method and Computer Program Product For Identifying Chemical Compounds Having Desired Properties."

Molecular Similarity. Similarity is one of the most subjective concepts in chemistry, and can be defined in a multitude of ways. Depending on one's objectives, available tools, and other factors, compounds can be considered similar if they have similar numbers of atoms of the same types (constitutional similarity), similar numbers of bonds and rings of the same types and similar degree of branching (topological similarity), similar shape and surface characteristics (shape similarity) or similar electron density distribution (electrostatic similarity). Alternatively, similarity can be determined based on the presence or absence of certain features such as a common substructure (substructural similarity), the relative position and orientation of important pharmacophoric groups (pharmacophore similarity), binding affinity as predicted by a receptor binding model (receptor affinity similarity), the degree of conformational overlap with a known receptor binder (conformational similarity), etc.

The precise numerical value used to describe the similarity between two compounds depends on the representation of these compounds, the weighting scheme used to scale different aspects of the representation, and the similarity coefficient used to compare these representations (Willett *et al.* "Chemical Similarity Searching," J. Chem. Inf. Comput. Sci. 1998, 38, 983-996). Often, individual compounds are represented by a bit-string, such as a substructure key or a hashed fingerprint, where each bit or group of bits indicates the presence or absence of a particular structural feature. Alternatively, compounds can be represented by a vector of real numbers, each of which corresponds to a particular molecular descriptor. It has been suggested that in all cases the representation of the structures must comply with the "neighborhood principle" if it is to be useful in identifying biologically active molecules (Cramer, *et al.* "Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research," J. Chem. Inf. Comput. Sci. 1998, 38, 1010-1023). The neighborhood principle states that molecules with similar representations (i.e. molecules located within the same local region or "neighborhood" of the feature space) should have similar values of the physical property of interest. Recently, the neighborhood

behavior of 11 sets of 2D and 3D molecular descriptors was analyzed following a validation study on the ability of these descriptors to cluster active compounds (Patterson *et al.* "Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors," J. Med.Chem. 1996, 39, 3049-3059; Brown *et al.* "Designing Combinatorial Library Mixtures Using a Genetic Algorithm," J. Med. Chem. 1997, 40, 2304-2313). Descriptors which were found to exhibit "proper neighborhood behavior" included 2D fingerprints, topomeric fields, and atom pairs. Finally, the degree of structural similarity between two compounds is quantified by means of a similarity coefficient, such as Tanimoto coefficient for binary sets, and the Euclidean distance for real vectors. A thorough review of molecular similarity measures can be found in Willett *et al.* "Chemical Similarity Searching," J. Chem. Inf. Comput. Sci. 1998, 38, 983-996, which is incorporated herein by reference in its entirety.

The specific details of the present invention are described in detail below, in relation to exemplary embodiments.

2. *Exemplary Embodiments*

As discussed above, the present invention is directed to the searching of large virtual combinatorial libraries based on a fitness function, such as similarity to a query structure. A preferred method for fast searching of large combinatorial libraries, according to the present invention, is described with reference to FIG. 1.

First, in step 104, a sufficiently large sample N (e.g., N = 100,000) of potential compounds (i.e., reagent combinations) is selected from a virtual combinatorial library. The selected sample is also referred to as the first set of reagent combinations. In a preferred embodiment, the set of N reagent combinations are selected at random. In such an embodiment, the sample is also referred to as the random set, the original random set, and the random seed. In another embodiment, the first set of reagent combinations is selected such that a

uniform coverage of reagent space is selected. Any known algorithm for selecting a random or approximately random set can be employed. In still another embodiment, the first set of reagent combinations is selected such the each reagent in the virtual combinatorial library is selected exactly a predefined (e.g., X) number of times or a substantially equal number of times. Of course other methods of selecting N reagent combinations are within the spirit and scope of the present invention, including random and non-random methods.

The selection of a random sample can be better explained using the following example. Assume a random sample of 100,000 possible reagent combinations (i.e., N = 100,000) are to be selected from a virtual combinatorial library that represents the chemical reaction $A_i + B_j + C_k \rightarrow D_{ijk}$, where reagent class A includes 1,000 reagents, reagent class B includes 10,000 reagents, and reagent class C includes 5,000 reagents. The virtual combinatorial library representing this chemical reaction can be thought of as a 1,000 x 10,000 x 500 matrix (i.e., a three dimensional matrix), with 5 x 10^9 possible products (i.e., reagent combinations). More specifically, along the X axis of the matrix is A_1 to $A_{1,000}$, along the Y axis of the matrix is B_1 to $B_{10,000}$, and along the Z axis of the matrix is C_1 to C_{500}. One method of selecting 100,000 possible combinations of reagents from this virtual combinatorial library is by generating 100,000 random numbers from 1 to 1000, 100,000 random numbers from 1 to 10,000, and 100,000 random numbers from 1 to 500, and using these randomly selected numbers to generate 100,000 possible combinations of A_i, B_j and C_k. Examples of possible combinations of A_i, B_j and C_k include: $A_{322}B_{1902}C_{401}$ (i.e., $A_{i=322}B_{j=1902}C_{k=401}$), $A_{332}B_{205}C_5$, $A_{105}B_{9333}C_{304}$ and $A_{46}B_{9502}C_{208}$. Each of these 100,000 possible combinations of A_i, B_j and C_k can be thought of representing Cartesian coordinates (i.e., empty cells of the matrix) corresponding to a particular combination of reagents. Of course, other methods of selecting a set are within the spirit and scope of the present invention.

Selection of an appropriate value N is implementation specific. As discussed in the Experimental Results section below, good results have been

obtained when the value chosen for N is approximately 0.1% of the total number R of reagent combinations associated with the virtual combinatorial library.

Next, in step 106, each reagent combination in the first set of compounds is enumerated to produce a first set of enumerated compounds. Thus, if 100,000 reagent combination (e.g., N = 100,000) were selected in step 104, then the first set of enumerated compounds will include 100,000 enumerated compounds.

Next, in step 108, M compounds (e.g., M = 100) are selected, based on a fitness function, from the first set of enumerated compounds.

Selection of an appropriate value M is implementation specific. As discussed in the Experimental Results section below, good results have been obtained when the value chosen for M is approximately 0.1% of the value N.

Once M compounds are selected, based on the fitness function, from the first set of enumerated compounds, these M compounds are then deconvoluted into their building blocks (i.e., reagents), in step 110.

In step 112, the building blocks resulting from step 110 are combined into lists of "preferred" reagents and are used to produce a smaller "focused" library. This "focused" library can be thought of as a sub-matrix of the larger matrix that represents the entire original virtual combinatorial library.

Substantially all of the potential compounds in the "focused" library (i.e., the second set of compounds) are then enumerated in step 114 to produce a second set of enumerated compounds. Since the focused library is significantly smaller than the original virtual combinatorial library, the amount of time and resources required to enumerate the entire focused library is significantly less than required to enumerate the entire original virtual combinatorial library.

Finally, in step 116, K compounds (e.g., K = 100) are selected, based on the fitness function, from the second set of enumerated compounds. These K compounds represent near optimal set of compounds that best satisfy the fitness function. The selection of an appropriate value K is also implementation specific.

Because of its stochastic nature, the best results are obtained by repeating the above described method more than one time and combining the results. The

inventors have found that repeating the above method (i.e., steps 104 - 116) three separate times and combining the results produces excellent results, as explained below, although the invention is not limited to this example.

The preferred method for fast searching of a large combinatorial library described above in the discussion of FIG. 1 can be further explained with reference to FIGS. 2, 3 and 4.

FIG. 2 shows a 4 x 4 matrix 201 of reagents that represents a small portion of a large virtual combinatorial library associated with the chemical reaction $A_i + B_j \rightarrow C_{ij}$. If the entire virtual combinatorial library were enumerated, then each of the 16 cells shown in the 4 x 4 matrix 201 would include an enumerated compound (i.e., a computer representation of the structure of a compound). As discussed above, a computer system may require an inordinate amount of time and resources to enumerate every potential reagent combination associated with a large virtual combinatorial library. Accordingly, in the present invention, a sample N of potential compounds (i.e., reagent combinations) is selected (e.g., at random) from the virtual combinatorial library (Step 104). In FIG. 2, the four cells, 202, 204, 206, and 208, that have thick borders represent the selected reagent combinations. As shown in FIG. 2, only those selected reagent combinations are enumerated (Step 106) to produce a first set of enumerated compounds which includes compounds 203, 205, 207, and 209. Next, M compounds are selected, based on the fitness function, from the first set of enumerated compounds (Step 108). For this example, it is assumed that enumerated compounds 203 and 205 are included in the M selected compounds.

Once M compounds are selected, based on the fitness function, from the first set of enumerated compounds, these M compounds are then deconvoluted into their building blocks (i.e., reagents) (Step 110). This is represented by the arrows in FIG. 3.

The building blocks are then combined into lists of "preferred" reagents that are used to produce a smaller "focused" library, which includes 2 x 2 matrix 401 (Step 112), as shown in FIG. 4. All the compounds in the "focused" library

are enumerated to produce a second set of enumerated compounds (Step 114), which includes compounds 403, 407, 409, and 405.

Finally, K compounds are selected, based on the fitness function, from the second set of enumerated compounds (Step 116). For this example, it is assumed that only enumerated compound 409 is included in the K selected compounds.

Another embodiment of the present invention takes advantage of those virtual libraries that have already gone through the timely process of being enumerated. More specifically, in an alternative embodiment, the present invention is directed to the searching of fully (or partially) enumerated combinatorial libraries based on a fitness function, such as similarity to one or more query structures. This method for fast searching of enumerated combinatorial libraries, according to this alternative embodiment of the present invention, is described with reference to FIG. 1A.

First, in step 104a, a sufficiently large sample N (e.g., N = 100,000) of enumerated compounds is selected from an enumerated virtual combinatorial library. The selected sample is also referred to as the first set of enumerated compounds. In one embodiment, the set of N enumerated compounds are selected at random. In another embodiment, the first set of N enumerated compounds is selected such that a uniform coverage of reagent space is selected. Any known algorithm for selecting a random or approximately random set can be employed. In still another embodiment, the first set of enumerated compounds is selected such the each reagent in the virtual combinatorial library is selected exactly a predefined (e.g., X) number of times or a substantially equal number of times. Of course other methods of selecting N enumerated compounds are within the spirit and scope of the present invention, including random and non-random methods.

The selection of a random set of enumerated compounds can be better illustrated using the following example. Assume a random selection of 100,000 enumerated compounds (i.e., N = 100,000) are to be selected from an enumerated virtual combinatorial library that represents the chemical reaction $A_i + B_j + C_k \rightarrow D_{ijk}$, where reagent class A includes 1,000 reagents, reagent class B includes

10,000 reagents, and reagent class C includes 5,000 reagents. The enumerated virtual combinatorial library representing this chemical reaction can be thought of as a 1,000 x 10,000 x 500 matrix (i.e., a three dimensional matrix), with 5×10^9 enumerated compounds within in the cells of the matrix. More specifically, along the X axis of the matrix is A_1 to $A_{1,000}$, along the Y axis of the matrix is B_1 to $B_{10,000}$, along the Z axis of the matrix is C_1 to C_{500}, and within the cells are enumerated compounds. One method of selecting 100,000 enumerated compounds from this enumerated virtual combinatorial library is by generating 100,000 random numbers from 1 to 1000, 100,000 random numbers from 1 to 10,000, and 100,000 random numbers from 1 to 500, and using these randomly selected numbers to select 100,000 enumerated compounds. Each of these 100,000 possible combinations of A_i, B_j and C_k can be thought of representing cells of the matrix that include the corresponding enumerated compound. Of course, other methods of selecting a set are within the spirit and scope of the present invention.

Selection of an appropriate value N is implementation specific. Good results have been obtained when the value chosen for N is approximately 0.1% of the total number enumerated compounds associated with the enumerated virtual combinatorial library.

Next, in step 108a, M compounds (e.g., M = 100) are selected, based on a fitness function, from the first set of enumerated compounds. Selection of an appropriate value M is implementation specific. Good results have been obtained when the value chosen for M is approximately 0.1% of the value N.

Once M compounds are selected, based on the fitness function, from the first set of enumerated compounds, these M compounds are then deconvoluted into their building blocks (i.e., reagents), in step 110a.

In step 112a, the building blocks resulting from step 110a are combined into lists of "preferred" reagents and are used to extract a smaller "focused" enumerated library from the enumerated virtual combinatorial library. This "focused" enumerated library, which can be thought of as a sub-matrix of the

larger matrix that represents the entire original enumerated virtual combinatorial library, is also referred to as the second set of enumerated compounds.

Finally, in step 116a, K compounds (e.g., K = 100) are selected, based on the fitness function, from the second set of enumerated compounds. These K compounds represent near optimal set of compounds that best satisfy the fitness function. The selection of an appropriate value K is also implementation specific.

Because of its stochastic nature, the best results are obtained by repeating the above described method more than one time and combining the results. The inventors have found that repeating the above method (i.e., steps 104a - 116a) three separate times and combining the results produces excellent results, although the invention is not limited to this example.

a. K Most Similar Compounds to a Query Structure

As mentioned above, the present invention can be used to efficiently and effectively perform similarity searching of a large virtual combinatorial library. That is, in one embodiment, the fitness function of steps 108 and 116 relates to molecular similarity.

In an example embodiment, where the fitness function is similarity to one or more query structures (i.e., lead(s) or probe(s)), step 108 of FIG. 1 can be broken into more detailed steps 502, 504, 506 and 508, and step 116 can be broken into steps 510, 512, 514 and 516, as described below, and as shown in FIG. 5.

As described above, in the discussion of FIG. 1, in steps 104 and 106, N reagent combinations are selected (e.g., randomly) from the virtual combinatorial library and each reagent combination is enumerated to produce a first set of enumerated compounds.

In step 502, the first set of enumerated compounds are characterized by calculating a prescribed set of molecular descriptors. These same descriptors are also calculated for the query structure(s) (also referred to as the probe(s)), which

can be, for example, one or more drug leads. The following articles, which are incorporated herein by reference in their entirety, describe suitable example molecular descriptors: Agrafiotis "On the Use of Information Theory for Assessing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 576-580; Agrafiotis, "Stochastic Algorithms for Maximizing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 841-851; and Agrafiotis, "Diversity of Chemical Libraries," Encyclopedia of Computational Chemistry 1998.

In step 504, the pairwise similarities between the query structure(s) and the first set of enumerated compounds (also referred to as, the first combinatorial compound set) are evaluated using a similarity measure of choice. For example, compounds can be considered similar if they have similar numbers of atoms of the same types (constitutional similarity), similar numbers of bonds and rings of the same types and similar degree of branching (topological similarity), similar shape and surface characteristics (shape similarity) or similar electron density distribution (electrostatic similarity). Alternatively, similarity can be determined based on the presence or absence of certain features such as a common substructure (substructural similarity), the relative position and orientation of important pharmacophoric groups (pharmacophore similarity), binding affinity as predicted by a receptor binding model (receptor affinity similarity), or the degree of conformational overlap with a known receptor binder (conformational similarity). The specific similarity measure chosen will depend on the objectives, available tools, and other factors.

Once the pairwise similarities between the query structure(s) and the first set of enumerated compounds are determined, the compounds within the first combinatorial compound set (the first set of enumerated compounds) are then preferably sorted in descending (or ascending) order of similarity to the probe, in step 506. The top-ranking (or bottom-ranking) M compounds, also referred to as the highest-ranking M compounds (or lowest-ranking M compounds), are then selected in step 508. Alternatively, the top-ranking (or bottom-ranking) M

compounds can be selected without prior sorting. For example, any compound having a dissimilarity value lower than a threshold value can be selected.

Next, as described above in the discussion of FIG. 1, in steps 110, 112, and 114, the M compounds are deconvoluted into reagents, a focused library is created, and all the reagent combinations in the focused library are enumerated to produce a second set of enumerated compounds.

In step 510, the compounds in the second set of enumerated compounds are characterized by calculating descriptors, preferably using the same set of molecular descriptors used to characterize the first set of enumerated compounds (in step 502) and the query structure(s).

Using the same query structure(s) (i.e., probe(s)) and the same similarity measure used in step 504 to screen the first set of enumerated compounds, the pairwise similarities between the query structure(s) and the second set of enumerated compounds are evaluated in step 512. These enumerated compounds are then preferably sorted in descending (or ascending) order of similarity to the probe(s) (i.e., query structure(s)) in step 514. Finally, in step 516 the desired number (e.g., K) of the highest-ranking (most similar) compounds are selected from the enumerated "focused" library (i.e., second set of enumerated compounds) based on their similarity scores or dissimilarity scores. Alternatively, the highest-ranking K compounds can be selected without prior sorting. For example, any compound having a dissimilarity value lower than a threshold value can be selected.

These K most similar compounds selected in step 516 can then be physically synthesized and screened to see if they include, for example, biologically active compounds or hits.

The present invention can also be used to efficiently and effectively perform similarity searching of a large enumerated virtual combinatorial library. That is, in one embodiment, the fitness function of steps 108a and 116a relates to molecular similarity. Where the fitness function is similarity to one or more query structures, steps 108a and 116a can be broken into more detailed steps. The more

detailed steps associated with step 108a are similar to steps 502-508, which are discussed above in the description of FIG. 5. The more detailed steps associated with step 116a are similar to the detailed steps 510-516, which are also discussed above in the description of FIG. 5.

As described above, in the discussion of FIG. 1A, in step 104a, enumerated compounds are selected (e.g., randomly) from an enumerated combinatorial library to produce a first set of enumerated compounds. Then in step 108a, M compounds are selected based on a fitness function. Where the fitness function is related to similarity to a query structure, this can be accomplished as described immediately below.

The first set of enumerated compounds can be characterized by calculating a prescribed set of molecular descriptors. The same descriptors are also calculated for the one or more query structures. Next, the pairwise similarities between the query structure(s) and the first set of enumerated compounds are evaluated using a similarity measure of choice. Once the pairwise similarities between the query structure(s) and the first set of enumerated compounds are determined, the compounds within the first set of enumerated compounds are sorted based on their similarity to the probe(s). The top (or bottom) ranking M compounds are then selected. Alternatively, the top (or bottom) ranking compounds are selected without prior sorting.

Next, as described above in the discussion of FIG. 1A, in steps 110a and 112a, the M compounds are deconvoluted into reagents, and an enumerated focused library (also referred to as a second set of enumerated compounds) is extracted from the enumerated virtual combinatorial library.

As described above, in step 116a, K compounds are selected based on the fitness function. Where the fitness function is related to similarity to a query structure, this can be accomplished as described immediately below.

The compounds in the second set of enumerated compounds are characterized by calculating descriptors, preferably using the same set of molecular descriptors used to characterize the first set of enumerated compounds and the

query structure(s). Then, using the same query structure(s) (i.e., probe(s)) and the same similarity measure used to screen the first set of enumerated compounds, the pairwise similarities between the query structure(s) and the second set of enumerated compounds are evaluated. These enumerated compounds are then sorted based on their similarity to the probe(s), and the desired number (e.g., K) of the highest (or lowest) ranking compounds is extracted from the second set of enumerated compounds based on their similarity scores or dissimilarity scores. Alternatively, the highest (or lowest) ranking K compounds can be selected without prior sorting. These K most similar selected compounds can then be physically synthesized and screened to see if they include, for example, biologically active compounds or hits.

b. Array (Sub-Matrix) of Most Similar Compounds

Similarity selections from virtual combinatorial libraries are aimed at producing candidates for future synthesis and biological testing. That is, the K compounds selected in step 516 can be synthesized and screened to identify "hits". In order to simplify and reduce the cost of synthesis, combinatorial compounds are typically synthesized in an array (i.e., sub-matrix) format. The stochastic procedure of the present invention described above can be adapted to generate such arrays (i.e., sub-matrices). After the "preferred" reagents have been identified and the focused library has been enumerated, a simulated annealing or genetic algorithm based search engine can be used to find a sub-matrix that exhibits the lowest average dissimilarity score (or highest average similarity score). In more general terms, some satisfaction of the fitness function (e.g., similarity to the probe structure) can be sacrificed in exchange for reducing the number T of reagents that make up the selected K compounds. An example of how this can be done is explained below.

Assume a virtual combinatorial library represents the chemical reaction $A_i + B_j \rightarrow C_{ij}$, where reagent class A includes 100,000 regents and reagent class B

also includes 100,000 reagents. Thus, the virtual combinatorial library representing this chemical reaction can be thought of as a 100,000 x 100,000 matrix, which includes 1×10^{10} reagent combinations (i.e., potential compounds). If the present invention were used to select the 100 most similar compounds in step 116 (i.e., K = 100), there is the possibility that 200 different reagents (i.e., 100 different A reagents and 100 different B reagents) would be required to synthesize the 100 most similar compounds. This can be costly and time consuming. Thus, it may be beneficial to choose 100 similar compounds that can be produced using, for example, only 20 different reagents (e.g., 10 different A reagents and 10 different B reagents). This can be accomplished, for example, by modifying the method of FIG. 5 so that steps 512, 514 and 516 are replaced with 612, 614 and 616, as shown in an embodiment of FIG. 6 and as described below.

In step 612, a sub-matrix of K compounds is selected, wherein the K compounds include a total number T of reagents. For example, if T = 20, then the sub-matrix can be a 10 x 10 (i.e., 10 A reagents and 10 B reagents), a 9 x 11, a 8 x 12, ... , a 2x18 or a 1 x 20 matrix.

In step 614, the similarity of the K compounds to the query structure(s) is evaluated using the same similarity measure used in step 504 to screen the original set. Preferably, this is accomplished by evaluating pairwise similarities between the query structure(s) and each compound in the sub-matrix. The result of step 614 can be a total similarity value or an average similarity value for the K compounds in the sub-matrix.

Next, in step 616, the set of K compounds is gradually refined by a series of small stochastic 'steps' to optimize the similarity of the sub-matrix. Here, the term 'step' is taken to imply a stochastic (random or semi-random) modification of the sub-matrix of K compounds. For example, the sub-matrix of K compounds can be modified by removing one of reagents associated with the sub-matrix and replacing it with a randomly selected reagent associated with the focused library. Alternatively, the sub-matrix of K compounds can be modified by altering the structure of the matrix, while keeping the total number T of reagents the same and

the number K of compounds the same. For example, the sub-matrix may be changed from a 10 x 10 matrix to a 11 x 9 matrix. After the 'step' is performed, the similarity of the resulting new set (i.e., sub-matrix) of K compounds is assessed, and the similarity of the new sub-matrix is compared to the similarity of the old sub-matrix. If the new sub-matrix has a greater similarity than the old set, it replaces the old sub-matrix and the process is repeated. If the new sub-matrix does not have a greater similarity than the old sub-matrix, then the old sub-matrix is retained (as the most similar sub-matrix containing K compounds made of T reagents). This process is repeated, until a sufficiently similar sub-matrix containing K combinatorial products is selected.

This process can be controlled, for example, by a Monte-Carlo sampling protocol, a Simulated Annealing protocol, or variants thereof, which are well known to people skilled in the art. However, it should be understood that the present invention is not limited to these embodiments. Alternatively, any other suitable search/optimization algorithm can be used. The implementation of these methods should be straightforward to persons skilled in the art.

Additionally, where the fitness function is related to similarity to a query structure(s), the number or reagents used to synthesize the K compounds selected from an enumerated virtual combinatorial library, in step 116a, can be reduced in a manner similar to that described above in the discussion of steps 612-616.

c. K Most Diverse Compounds

The present invention can also be used to efficiently and effectively perform dissimilarity (diversity) searching of a large virtual combinatorial library. Accordingly, in one embodiment, the fitness function referred to in step 108 and 116 relates to molecular dissimilarity or diversity. In such an embodiment, the present invention can be used to identify the most diverse set of K compounds of a large virtual combinatorial library.

In contrast to similarity, which is typically a pairwise fitness function characterizing the degree of likeness between a given compound and a probe or target structure, diversity is a fitness function associated with a collection of compounds. The evaluation of the diversity of a collection of compounds generally involves two steps: 1) evaluation of the pairwise dissimilarities of the compounds, and 2) evaluation of the diversity of the collection based on these pairwise dissimilarities. Evaluation of pairwise dissimilarities is similar to the evaluation of pairwise similarities described above, with the exception that there is no probe structure. Hence, any suitable similarity metric and choice of descriptors can be applied. Finally, a diversity metric is used to evaluate the diversity of a collection of compounds. The following articles, which have been incorporated herein by reference in their entirety, describe such suitable similarity metric and descriptors, although the invention is not limited to that described therein: Agrafiotis "On the Use of Information Theory for Assessing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 576-580; Agrafiotis, "Stochastic Algorithms for Maximizing Molecular Diversity," J. Chem. Inf. Comput. Sci. 1997, 37, 841-851; and Agrafiotis, "Diversity of Chemical Libraries," Encyclopedia of Computational Chemistry 1998.

Since the evaluation of diversity typically involves pairwise evaluation of dissimilarities, the number of dissimilarities which have to be evaluated grows exponentially with the size of the collection. Due to this exponential increase in computational complexity, selecting a most diverse set of compounds can be a very challenging task. The present invention provides an efficient and effective solution for selecting a diverse set of compounds from a large virtual combinatorial library.

In a preferred implementation, a stochastic version of a maximin selection algorithm is used to select a diverse subset of M compounds from the N combinatorial products selected at random. In the maximin algorithm, the diversity of a collection of compounds, can be, for example, evaluated as the minimal pairwise distance (or minimal pairwise dissimilarity) between any two

compounds in the collection. A simulated annealing stochastic procedure can then be used to modify the collection with a goal to maximize the minimal pairwise distance (hence maximin).

The well known maximin function that is referred to above is:

$$f(S) = \max_{i}(\min_{j \neq i}(d_{ij}))$$

or its variant:

$$f(S) = \sum_{i}(\min_{j \neq i}(d_{ij}))$$

where

S is any given M-membered subset of the N-membered virtual combinatorial library; and i, j are used to index the elements of S.

This function has the advantage that it can be used with any conceivable dissimilarity index and does not require a metric space. In practice, the inventors have found the later equation to be smoother and thus much easier to optimize in a Monte-Carlo environment. Of course any other suitable selection algorithm can be used to select a diverse subset of M compounds from the N combinatorial products.

More specifically, where the fitness function is diversity, step 108 of FIG. 1 can be broken into more detailed steps 702, 704, 706 and 708, and step 116 can be broken into steps 710, 712, 714 and 716, as shown in the example embodiment of FIG. 7 and described below.

As described above, in the discussion of FIG. 1, in steps 104 and 106, N reagent combinations are selected (e.g., randomly) from the virtual combinatorial library and each reagent combination is enumerated to produce a first set of enumerate compounds.

In step 702, the first set of N enumerated compounds are characterized by calculating a prescribed set of molecular descriptors. In step 704, an initial set of M compounds is selected (e.g., at random) from the first set of N enumerated

compounds. The diversity of the M compounds is then evaluated using methods known in the art in step 706. As discussed above, the evaluation of the diversity of a collection of compounds generally involves evaluating the pairwise dissimilarities of the M compounds, and evaluating the diversity of the collection based on these pairwise dissimilarities.

Next, in step 708, the set of M compounds is gradually refined by a series of small stochastic 'steps' to optimize the diversity of the set. Here, the term 'step' is taken to imply a stochastic (random or semi-random) modification of the set of M compounds. For example, the set of M compounds can be modified by removing one of the M compounds and replacing it with a randomly selected one of the compounds in the first set of enumerated compounds. After the 'step' is performed, the diversity of the resulting new set of M compounds is assessed, and the diversity of the new set is compared to the diversity of the old set using any suitable comparison criterion. If the new set has a greater diversity than the old set, it replaces the old set and the process is repeated. If the new set does not have a greater diversity than the old set, then the old set is retained (as the most diverse set) and the process is repeated. This process is preferably repeated until the diversity of the set of M compounds essentially plateaus. This process can be controlled, for example, by a Monte-Carlo sampling protocol, a Simulated Annealing protocol, or variants thereof, which are well known to people skilled in the art. However, it should be understood that the present invention is not limited to these embodiments. Alternatively, any other suitable search/optimization algorithm can be used. The implementation of these methods should be straightforward to persons skilled in the art.

After a sufficiently diverse subset of M combinatorial products is selected, contributing reagents are identified in step 110 by deconvoluting the M compounds. A focused library is then created in step 112, and all the reagent combinations of the focused library are enumerated in step 114 to create a second set of enumerated compounds.

Finally, the selection algorithm is applied again to select a sufficiently diverse subset of K compounds from the focused library. More specifically, in step 710 molecular descriptors are calculated for each of the enumerated compounds. In step 712, an initial set of K compounds is selected (e.g., at random) from the second set of enumerated compounds. The diversity of the K compounds is then evaluated using methods known in the art in step 714. Next, in step 716, the set of K compounds is gradually refined by a series of small stochastic 'steps', as in step 708, until a nearly optimal diverse subset of K combinatorial products is selected.

An embodiment of the present invention can also be used to efficiently and effectively perform dissimilarity (diversity) searching of a large enumerated virtual combinatorial library. Accordingly, in one embodiment, the fitness function referred to in steps 108a and 116a relate to molecular dissimilarity or diversity. In such an embodiment, the present invention can be used to identify the most diverse set of K compounds of a large enumerated virtual combinatorial library.

More specifically, where the fitness function is diversity, steps 108a and 116a of FIG. 1A can be broken into more detailed steps. The more detailed steps corresponding to step 108a are similar to steps 702-708, which are discussed above in the description of FIG. 7. The more detailed steps corresponding to step 116a are similar to steps 710-716, which are also discussed above in the description of FIG. 7.

As described above, in the discussion of FIG. 1A, in step 104a, N enumerated compounds are selected (e.g., randomly) from an enumerated virtual combinatorial library to produce a first set of enumerate compounds. Where the fitness function is diversity, the method of selecting M compounds in step 108a can be accomplished as described immediately below.

First, the set of N enumerated compounds are characterized by calculating a prescribed set of molecular descriptors. Next, an initial set of M compounds is selected (e.g., at random) from the first set of N enumerated compounds. The diversity of the M compounds is then evaluated using methods known in the art.

As discussed above, the evaluation of the diversity of a collection of compounds generally involves evaluating the pairwise dissimilarities of the M compounds, and evaluating the diversity of the collection based on these pairwise dissimilarities.

Next, the set of M compounds is gradually refined by a series of small stochastic 'steps' to optimize the diversity of the set. Here, the term 'step' is taken to imply a stochastic (random or semi-random) modification of the set of M compounds. For example, the set of M compounds can be modified by removing one of the M compounds and replacing it with a randomly selected one of the compounds in the first set of enumerated compounds. After the 'step' is performed, the diversity of the resulting new set of M compounds is assessed, and the diversity of the new set is compared to the diversity of the old set using any suitable comparison criterion. If the new set has a greater diversity than the old set, it replaces the old set and the process is repeated. If the new set does not have a greater diversity than the old set, then the old set is retained (as the most diverse set) and the process is repeated. This process is preferably repeated until the diversity of the set of M compounds essentially plateaus. This process can be controlled by any suitable search/optimization algorithm/process, including, but not limited to, a Monte-Carlo sampling protocol, a Simulated Annealing protocol, or variants thereof, which are well known to people skilled in the art.

After a sufficiently diverse subset of M enumerated compounds is selected, contributing reagents are identified by deconvoluting the M compounds. An enumerated focused library is then extracted, based on the reagents, from the enumerated virtual combinatorial library. The enumerated focused library is also referred to as the second set of enumerated compounds.

Finally, the selection algorithm is applied again to select a sufficiently diverse subset of K compounds from the enumerated focused library. More specifically, molecular descriptors are calculated for each of the enumerated compounds in the enumerated focused library. An initial set of K compounds is selected (e.g., at random) from the second set of enumerated compounds and the diversity of the K compounds is then evaluated. The set of K compounds is

gradually refined by a series of small stochastic 'steps' until a nearly optimal diverse subset of K enumerated compounds is selected.

d. Array (Sub-Matrix) of Most Diverse Compounds

Diversity selections from virtual combinatorial libraries are also aimed at producing candidates for future synthesis and biological testing. That is, the K compounds selected in step 716 can be synthesized and screened to identify "hits". In order to simplify and reduce the cost of synthesis, combinatorial compounds are typically synthesized in an array (i.e., sub-matrix) format. The stochastic procedure of the present invention can be easily adapted to generate such arrays (i.e., sub-matrices). That is, some satisfaction of the fitness function (e.g., dissimilarity of a collection of compounds) can be sacrificed in exchange for reducing the total number T of reagents that make up the K compounds. This can be accomplished, for example, by modifying the method of FIG. 7 so that steps 712 and 716 are replaced with steps 812 and 816, as shown in FIG. 8 and as described below.

In step 812, a sub-matrix of K compounds is selected, such that the K compounds include a total of T reagents. In step 714, the dissimilarity of the collection of K compounds is evaluated. Next, in step 816, the set (i.e., sub-matrix) of K compounds is gradually refined by a series of small stochastic 'steps' to optimize the dissimilarity of the sub-matrix. For example, the sub-matrix of K compounds can be modified by removing one of reagents associated with the sub-matrix and replacing it with a randomly selected reagent associated with the focused library. Alternatively, the sub-matrix of K compounds can be modified by altering the structure of the sub-matrix, while keeping the total number T of reagents the same and the number K of compounds the same. For example, the sub-matrix may be changed from a 10 x 10 matrix to a 11 x 9 matrix. After the 'step' is performed, the diversity (i.e., dissimilarity) of the resulting new set of K compounds is assessed, and the diversity of the new sub-matrix is compared to the

diversity of the old sub-matrix. If the new sub-matrix has a greater diversity than the old set, it replaces the old sub-matrix and the process is repeated. If the new sub-matrix does not have a greater diversity than the old sub-matrix, then the old sub-matrix is retained (as the most dissimilar sub-matrix containing K compounds made of T reagents). This process is repeated, until a sufficiently dissimilar sub-matrix containing K combinatorial products is selected. For example, this process can be repeated until the diversity of the K compounds essentially plateaus.

This process can be controlled, for example, by a Monte-Carlo sampling protocol, a Simulated Annealing protocol, or variants thereof, which are well known to people skilled in the art. However, it should be understood that the present invention is not limited to these embodiments. Alternatively, any other suitable search/optimization algorithm can be used. The implementation of these methods should be straightforward to persons skilled in the art.

Additionally, where the fitness function is related to dissimilarity (diversity) searching of a large enumerated virtual combinatorial library, the number of reagents used to synthesize the K compounds selected from the enumerated virtual combinatorial library, in step 116a, can be reduced in a manner similar to that discussed above in the description of steps 710, 812, 714 and 816.

For additional details of such an example simulated annealing or genetic search engine refer to the article entitled "Stochastic Algorithms for Maximizing Molecular Diversity," which has been incorporated by reference above.

3. Experimental Results and Discussion

The effectiveness of the present invention has been demonstrated by the inventors through experiments using two different virtual combinatorial libraries. The first was a diamine virtual combinatorial library, generated by combining a diamine core with a set of alkyl halides or acid chlorides, as shown in FIG. 9a. The structure 902a represents a diamine; R1-X and R2-X each independently represent an alkyl halide or acid chloride. Although physical synthesis of this

library could prove problematic (the synthetic sequence involves selective protection of one of the amines and introduction of the first side chain, followed by deprotection and introduction of the second side chain), for the purpose of a study it can be assumed that one of the amino groups on the diamine core reacts with the first reagent, while the other reacts with the second reagent. A substructure search in the Available Chemicals Directory (ACD) yielded 1,036 suitable diamines and 826 alkylating/acylating agents. These reagents were used to generate a virtual combinatorial library associated with over 706 million (1036 × 826 × 826) possible products (i.e.,reagent combinations). Since descriptors for the fully enumerated virtual combinatorial library could not be computed in a timely fashion, and since for validation purposes the inventors needed to compare their results with conventional selections from a fully characterized library, a smaller 6.75 million-membered library (i.e., a virtual combinatorial library associated with 6.75 million reagent combinations) was produced by choosing 300 diamines and 150 alkylating/acylating agents at random. Hereafter, the term "diamine virtual combinatorial library" will refer to this smaller library, unless noted otherwise. Examples of diamines, and acidchlorides and halocarbons (i.e., alkylating/acylating agents), associated with this smaller library are shown in FIG. 21.

The second virtual combinatorial library was based on the Ugi reaction, and involves an organic acid (R1-COOH), an amine (R2-NH_2), an aldehyde (R3-CHO) and an isonitrile (R4-CN), as shown in FIG. 9b. A substructure search in the ACD yielded 1,681 suitable acids, 594 suitable amines, 37 suitable aldehydes, and 17 suitable isonitriles. These reagents were used to build a virtual combinatorial library associated with over 628 million possible compounds (1681 × 594 × 37 × 17). Again, for validation purposes a smaller 6.29 million-membered library (i.e., a virtual combinatorial library associated with 6.29 million reagent combinations) was produced by choosing a random set of 100 acids and 100 amines. Hereafter, the term "Ugi virtual combinatorial library" will refer to this

smaller library, unless noted otherwise. Examples of acids, amines, aldehydes, and isonitriles associated with this smaller library are shown in FIG. 22.

First, every reagent combination (approximately 6.75 million) associated with the diamine virtual combinatorial library was enumerated to produce a full set of enumerated compounds. The similarities of the enumerated compounds to a query structure (an antiarrhythmic agent), shown in FIG. 10a, were then evaluated. The evaluation of molecular similarity was based on a standard set of 117 topological descriptors computed using a C++ descriptor generation class from the DirectedDiversity® API toolkit, available from 3-Dimensional Pharmaceuticals, Inc., Exton, Pennsylvania. The descriptors included a well-established set of topological indices with a long, successful history in structure-activity correlation such as molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstis indices, and topological state indices. These indices are discussed in detail in Hall *et al.* "The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Relations," Reviews of Computational Chemistry, Chap. 9, pp 367-422, eds. Donald Boyd and Ken Lipkowitz, VCH Publishers, Inc. (1991), and Bonchev *et al.* "Information Theory, Distance Matrix, and Molecular Branching," J. Chem. Phys. 1977, 67, pp 4517-4533, both of which are incorporated herein by reference in their entirety. The calculated descriptors were normalized, and decorrelated using principal component analysis. The principal components which accounted for 99% of the total variance in the data (typically 25-30 principal components) were used to define the similarity space. Pairwise dissimilarity scores were calculated as Euclidean distances between the vectors associated with the respective compounds in the space defined by the selected principal components. A higher dissimilarity score indicates compounds that are less similar to each other, that is, more distant from each other in the principal component space.

Based on calculated dissimilarity scores, a similarity profile 1102a, shown in FIG. 11a, of the diamine virtual combinatorial library was obtained by counting the number of compounds falling in each similarity bin 1104a. According to the

distribution, the majority of compounds had dissimilarity scores higher than 4.0. Next, the 100 most similar compounds with the lowest dissimilarity score were selected, and were used as a reference to compare all subsequent similarity selections drawn from the diamine virtual combinatorial library. These reference compounds represent the absolute best similarity selection of that size which can be obtained from the diamine virtual combinatorial library using the prescribed descriptors, similarity measure, and query structure.

Likewise, every reagent combination (approximately 6.29 million) associated with the Ugi virtual combinatorial library was also enumerated and the similarities of the enumerated compounds to the query structure (a 1.4 μM thrombin inhibitor), shown in FIG. 10b, were evaluated. Although the enumerated compounds associated with the Ugi virtual combinatorial library exhibited a more sharp similarity distribution, as shown by the similarity profile 1102b in FIG. 11b, the vast majority of the compounds in that library also had dissimilarity scores higher than 4.0. Again, the 100 most similar compounds were identified and were used as a reference to compare subsequent similarity selections from that library.

Next, the selection method of the present invention, outlined in the discussion of FIG. 5 above, was employed to select the 100 highest-scoring compounds from the diamine combinatorial virtual library based on their similarity to the same query structure of FIG. 10a. Initially 100,000 reagent combinations were selected at random from the virtual combinatorial library (Step 104, Fig. 5) and enumerated (Step 106). Descriptors were calculated for the enumerated compounds (Step 502), pairwise similarity to the query structure was evaluated (Step 504), and the enumerated compounds were ranked based on their similarity (Step 506). The 100 highest-ranking compounds were selected (508) and deconvoluted to produce lists of "preferred" reagents (Step 110). The list of "preferred" reagents was then used to produce the "focused" library (Step 112) and all reagent combinations associated with that "focused" library were enumerated (Step 114). Descriptors were calculated for the enumerated compounds (Step 510), pairwise similarity to the query structure was evaluated

(Step 512), and the enumerated compounds were ranked based on their similarity (Step 514). The 100 highest-ranking (most similar) compounds were then selected (Step 516) from the fully enumerated "focused" library based on their similarity to query structure of FIG. 10a

Because 100,000 reagent combinations of the virtual combinatorial library were enumerated (i.e., N = 100,000), and the 100 highest-ranking compounds were selected and deconvoluted to produce lists of "preferred" reagents (i.e., M = 100), the selection cycle was accordingly code-named 100K/100, and this naming scheme was used for all the remaining selections.

The dissimilarity scores and identities of the selected 100 compounds derived using the present invention were compared with the dissimilarity scores and identities of the reference selection derived from the fully enumerated library. Based on the average dissimilarity scores of the selected compounds (1.37 vs 1.30), the two selections (i.e., the selection using the present invention and the reference selection) were quite comparable. In fact, as shown in FIG. 12a, which shows the overlap between stochastic selections 1202a, 1204a, 1206a, and reference selection 1218a, most of the compounds in the reference set were also found in the stochastic selection. After repeating the stochastic procedure two more times with different random seeds and combining the results, the overlap with the reference selection rose to 96 out of 100 compounds, as shown at 1208a of FIG. 12a.

The same procedure was also applied to the Ugi library and an even better overlap with the reference selection was achieved, as shown in FIG. 12b.

Referring to FIG. 13, in three independent 100K/100 runs on the diamine virtual combinatorial library, the 100 highest-ranking compounds selected in Step 508 were derived, on average, from 108 reagents, which produced approximately a 50,000-membered enumerated focused library during Step 112, as shown in row 1304, of Table 1302. Thus, even though only approximately 450,000 compounds (or approximately 7% of the possible reagent combinations associated with the diamine virtual combinatorial library) were explicitly enumerated, described, and

compared to the query structure over all three runs, 96% of the best possible hits (i.e., most similar compounds) were retrieved. In fact, the number of unique compounds screened is substantially lower due to the substantial overlap between the focused libraries generated in the three independent runs.

For comparison, as shown in row 1306 of Table 1302, when three independent screens of 200,000 random reagent combinations relating to the diamine virtual combinatorial library were selected and enumerated, a total of only 9 (or 3 on average) out of the 100 most similar structures were retrieved. (This is also shown in FIG. 12a at 1210a, 1212a, 1214a and 1216a.) This result is not surprising since 600,000 compounds constitute approximately 10% of the entire 6.75 million-member diamine library.

In the case of the Ugi virtual combinatorial library, the 100 highest-ranking compounds selected during Step 508 produced an average of 99 "preferred" reagents. However, since this is a 4-component library, the resulting fully enumerated focused libraries were larger - 270,000 compounds on average. Thus, after three independent runs, a relatively higher proportion of the possible compounds associated with the Ugi virtual combinatorial library was screened (18%), which probably explains the higher percentage (98%) of compounds recovered from the reference set, as shown in row 1308 of Table 1302.

For comparison, as shown in row 1310 of Table 1302, when three independent screens of 400,000 random reagent combinations relating to the Ugi virtual combinatorial library were selected and enumerated, a total of only 17 (or 6 on average) out of the 100 most similar structures were retrieved.

The efficiency of the method of the present invention should be judged, for example, by two factors: 1) how good is the final selection, and 2) how many virtual compounds were actually screened (also referred to as cost). These two criteria are naturally connected: for example, if all virtual compounds were enumerated and screened, then the best possible selection would have been obtained. The goal is to find a nearly optimal set by enumerating and comparing the smallest possible number of compounds, and thus complete the task in a

reasonable time frame. The two parameters of the selection procedure that affect its outcome are the size N of the initial pool (Step 104) and the number M of the highest-ranking compounds (selected in Step 508) used to generate the focused library (in Step 112). In order to assess the effect of these parameters, a series of selections were carried out using several combinations of these parameters. For each combination, three independent runs were carried out starting from a different random seed, and the results were combined and summarized in Table 1302.

The number of highest-ranking compounds (i.e., the number M of compounds selected in step 508) used to determine the "preferred" reagents, which were used to produce the focused libraries, appears to have had the most significant effect on the quality of the final selection. This is not surprising, since a smaller number of compounds produces a smaller list of reagents which leads to a smaller focused library and, therefore, a smaller chance to retrieve the best hits. On the other hand, if that number is too large, much larger focused libraries are produced, and the execution speed of the algorithm is compromised, as shown in FIG. 14. Thus, the optimal number of highest-ranking compounds must be determined based on both the quality and cost of the final selection.

Graph 1402a shows the effect of the number of top-ranked M compounds chosen in step 508 on the quality and cost of the final selection from the diamine virtual combinatorial library. Graph 1402b shows the effect of the number of top-ranked compounds chosen in step 508 on the quality and cost of the final selection from the Ugi virtual combinatorial library. In each graph, the selection quality, shown as a solid line 1404, is measured in the percent overlap with the corresponding reference selection. The cost, shown as a dotted line 1406, is the cumulative percent of total number of virtual compounds evaluated. The cost is directly proportional to the amount of time it takes to perform enumeration.

For the diamine virtual combinatorial library, 100 was the optimal number, whereas for the Ugi library just 50 compounds were sufficient to obtain the best selection. Undoubtedly, the optimal number of compounds to choose will vary

from one virtual library to another, and will depend on the query structure as well. One can start with a small number and gradually increase it until the average dissimilarity score of the final selection reaches a plateau, as shown in FIG. 15. Specifically, FIG. 15 shows the average dissimilarity score of the final selection (selected in Step 516) as a function of the number M of top-ranked compounds chosen in step 508.

The inventors' experiments indicate that the size N of the initial pool of compounds selected at random (in Step 104) has a lesser effect on the quality of the final selection, as shown in graphs 1602a and 1602b of FIG. 16. However, if these initial pools are too small, they do not provide enough data to generate reliable statistics to determine the "preferred" reagents. If reagent combinations associated with a virtual combinatorial library are selected at random, they will sample all the reagents with the same probability. For example, if only 1,000 reagent combinations are selected from the 300x150x150 diamine virtual combinatorial library, every R1 reagent will be present in only 3 compounds on average, while every R2 and R3 reagent will be present in only 6 compounds on average. In fact, the probability to "miss" at least one reagent from that library is almost 1 if only 1,000 reagent combinations are selected, whereas that probability is almost 0 if 10,000 reagent combinations are selected.

If the initial pool is too small, each of the highest-ranking compounds can contribute completely different reagents, and the subsequent "focused" library can become too large and not focused at all. In this case, the similarity search degrades to a brute-force, random sampling approach, and becomes inefficient. On the other extreme, when the initial pool is too large, the number of "preferred" reagents and the resulting focused library decreases, but the cost (i.e., time) of the initial screening increases. In the inventors' experience, and as shown in FIG. 16, it is optimal to randomly select approximately 0.1% of the compounds in a virtual combinatorial library in the initial stage (Step 104) to achieve a nearly perfect similarity selection and at the same time keep the search practical.

Since the present invention is most useful when applied to massive virtual combinatorial libraries that are intractable by other means, the inventors derived a series of selections from the full diamine virtual combinatorial libraries and Ugi virtual combinatorial libraries containing 706 and 628 million possible enumerated compounds, respectively, using the same query structures of FIG. 10a and FIG. 10b, and varying the same selection parameters (i.e. the size N of the initial pool and the number M of highest-ranking compounds used to derive the focused library). As before, each combination of parameters was tested three times starting from a different random seed, and the results were averaged. The results are summarized in FIG. 17.

Referring to FIG. 17, graphs 1702a and 1704a show the dependence of the quality (i.e., average dissimilarity score) and cost (i.e., number of compound evaluated) of the final selection on the size N of the initial random pool of compounds (Step 104) for the full size diamine virtual combinatorial library. Similarly, graphs 1702b and 1704b show how dependence of the quality (i.e., average dissimilarity score) and cost (i.e., number of compound evaluated) of the final selection on the size N of the initial random pool of compounds (Step 104) for the full size Ugi virtual combinatorial library. Since the quality values relate to *dis*similarity scores, the lower the score the better (i.e., more similar) the result. That is, the lower the "Quality of the selection" value, the better the Quality of the selection.

As shown in these graphs, the quality of the combined selections (shown as solid lines 1706) is substantially better than the average quality of the individual selections (shown as the dashed lines 1708). Also shown is that with an increase of the size of the initial pool, the cost of the selection generally increases (as shown by the dash-dotted lines 1710), and the size of the focused library gets smaller (as shown by the dotted lines 1712).

It is clear that excellent selections were obtained in both cases after enumerating and screening on average less than 0.2% of the possible compounds in these virtual combinatorial libraries. In this case the inventors did not know (a

priori) the best possible hits, but the average dissimilarity score of the selections derived from the full libraries is substantially lower than that of the corresponding selections from the smaller libraries (0.2 vs. 1.3, and 0.7 vs. 1.4 for the diamine and Ugi libraries, respectively). Fig. 18 shows the structures of some of the most similar compounds found (in step 516). Structure 1802a, 1804a, and 1806a show three structures that were selected from the K most similar compounds associated with the diamine library. Structure 1802b, 1804b, and 1806b show three structures that were selected from the K most similar compounds associated with the Ugi library. An important finding is that in every case the quality of a combined selection after three independent runs (solid lines 1706 in Fig. 17) was substantially better than each individual selection (dashed lines 1708). In fact, the inventors found that better selections can be obtained by using a smaller initial pool and repeating the selection process three times, than by running the selection once from a three times larger pool, both in terms of quality and speed.

These last selections also confirmed that for the diamine library choosing the 100 instead of 50 highest-ranking compounds leads to a better selection, whereas for the Ugi library the improvement in quality is marginal, and is outweighed by the increase in cost (dash-dotted lines 1710). However, as pointed out earlier, the two parameters affecting the outcome of the selection procedure are not entirely independent. When the initial pool is large and the number of highest-ranking compounds selected is small, the resulting focused library is small. In fact, the larger the initial pool, the smaller the focused library (dotted lines 1712). Since the number M of highest-ranking compounds chosen to generate the focused library remains constant, this means that these compounds contribute fewer reagents. A likely consequence of this is the possibility of missing some of the most similar compounds. Imagine, for example, that the single most similar compound in a virtual library is built from two reagents, the first of which, *A*, represents 70% of the product's structure and the second, *B*, the remaining 30%. Since the initial compounds are selected at random, it is very unlikely that the product *AB* will be picked. However, there will be several compounds containing

only A (*A*-compounds) and several compounds containing only B (*B*-compounds). The larger the number of compounds screened, the more *A*- and *B*-compounds will be sampled. Because reagent A represents 70% of the product, the similarity of the *A*-compounds will be higher than that of the *B*-compounds. Therefore, if only a small number M of the highest-ranking compounds are chosen to generate the focused library, these will be exclusively *A*-compounds and the best compound *AB* will be never discovered. This effect can be seen in FIG. 17 where the quality value (i.e., dissimilarity) of the selection begins to decrease as the size of the initial pool becomes larger. Thus, a larger number M of highest-ranking compounds should be considered in such a case.

An important quality of the algorithm presented herein is the ability to produce very good hit lists in a short period of time. When less than 1% of the virtual compounds need to be enumerated and characterized, the effective performance gain is 100 fold. Additional performance enhancements can be achieved by enumerating the compounds and calculating descriptors in parallel on multiple CPU's. For example, the enumeration and similarity evaluation of all 6.75 million compounds in the diamine library required 34 hours on a dual processor 400MHz Pentium II machine. The stochastic algorithm using an initial pool of 100,000 and a focused library derived from the 100 highest-ranking compounds produced 88 of the 100 most similar compounds, and required only 30 minutes on the same system. A 1000K/100 selection from the 628 million-membered Ugi library takes less than 2 hours on a 6-processor R10,000 SGI. The inventors believe that this performance represents a dramatic improvement over conventional methodologies, and allows these methods to be used in a routine fashion.

4. Example Environment

An example environment in which the present invention is useful is described with reference to FIG. 19. As shown in FIG. 19, the example

environment includes virtual combinatorial libraries 1902, system and computer program product 1904, synthesis module 1906, and analysis module 1908. Virtual combinatorial library 1902 can be thought of as being essentially a computer representation of a collection of unenumerated chemical compounds. These computer representations are preferably stored in a database. System and computer program product 1904, can be, for example, a dual processor 400 MHZ Intel Pentium II machine running software. It is anticipated that the present invention is part of, or is performed by, system and computer program product 1904. As described below, the present invention can be used in this example environment to optimize the selection of lead compounds.

Initially, a relatively small sample of compounds are selected from the large virtual combinatorial library 1902. Preferably this sample of compounds is selected based on diversity criteria, and possibly considerations of cost, availability of reagents, ease of synthesis, and the like. Diversity is a very important selection criteria in selecting this initial sample of compounds because the sample should be as representative as possible of the entire virtual combinatorial library. The system and method of the present invention, described in FIGS. 7 and 8, can be used to select this diverse sample of compounds. Alternatively, other systems and methods can be used for selecting a diverse sample of compounds from a large combinatorial virtual library. Examples of such systems and methods are described in the article "Stochastic Algorithms for Maximizing Molecular Diversity," which has been incorporated by reference above.

Once a diverse sample of compounds is selected, each compound is physically synthesized to create a probe library 1910. The synthesized compounds of probe library 1908 are then screened by analysis module 1908. Analysis module 1908 preferably assays the synthesized compounds to obtain activity data, such as, enzyme activity data, cellular activity data, toxicology data, and/or bioavailability data. Analysis module 1908 can also analyze the compounds to obtain other pertinent data, such as structure and electronic structure data. Analysis module 1908 can then identify those compounds that possess the most desired properties

as lead compounds 1914 (also referred to as "hits"). This can be accomplished by assigning an activity value to each synthesized compound, and selecting those compounds with the highest activity values. Alternatively, lead compounds can be identified by any other method for lead compound identification. For example, the lead compounds may be identified in printed publications, traditional SAR studies, and prior combinatorial chemistry experiments.

In the embodiment where the present invention is used for fast similarity searching of a virtual combinatorial library 1902, these lead compounds (i.e., hits) are the query structures that are used in the present invention (i.e., in steps 504 and 512). When used in such an embodiment, the K (e.g., 100) most similar compounds 1912 to the query structure are obtained. These most similar compounds 1912 can in turn be synthesized by synthesis module 1906. Analysis module 1908 then screens and analyzes the synthesized compounds to obtain new lead compounds 1914. These lead compounds 1914 can then be used as new query structures for use with the present invention. That is, these new lead compounds can be used to generate a new selection of the K most similar compounds 1912. The above described process can be repeated again and again to create additional leads. In other words, the present invention can be used as part of an iterative process for generating lead compounds.

Recently, systems have been developed that permit the automatic chemical synthesis, refinement, and elaboration of bioactive compounds through the tight integration of high-speed parallel synthesis, structure-based design, and chemi-informatics. Such systems, known as DirectedDiversity systems, use an iterative optimization process that explores combinatorial space through successive rounds of selection, synthesis, and testing. Examples of these DirectedDiversity systems are described in the following patents and patent application, each of which are incorporated herein by reference in its entirety: U.S. Pat. No. 5,463,564, entitled "System and Method of Automatically Generating Chemical Compounds with Desired Properties"; U.S. Pat. No. 5,574,656 entitled "System and Method of Automatically Generating Chemical Compounds with Desired Properties"; U.S.

Pat. No. 5,684,711, entitled, "System, Method, and Computer Program for at Least Partially Automatically Generating Chemical Compounds Having Desired Properties"; U.S. Pat. No. 5,901,069, entitled "System, Method, and Computer Program Product for at Least Partially Automatically Generating Chemical Compounds with Desired Properties from a List of Potential Chemical Compounds to Synthesize"; and U.S. Pat. Appl. No. 08/963,870, entitled "System, Method and Computer Program Product For Identifying Chemical Compounds Having Desired Properties."

Unlike traditional combinatorial approaches where the entire library is made and tested in a single conceptual step, DirectedDiversity systems physically synthesize, characterize, and test only a portion of that library at a time. The selection of compounds is carried out by computational search engines that combine optimal exploration of molecular diversity with a directed search based on SAR information accumulated from previous iterations of the integrated machinery.

A central task of DirectedDiversity systems is to select an appropriate set of compounds for physical synthesis and biological evaluation. The present invention provides an efficient system and method for selecting such an appropriate set of compounds. That is, the present invention can be used in the systems described in the above listed patents and application, to generate the list of K compounds to be synthesized during each iteration.

5. Structure of Present Invention

It is anticipated that the present invention can be implemented as hardware, firmware, software or any combination thereof, and can be implemented in one or more computer systems and/or other processing systems. In one embodiment, the present invention is implemented by one or more computer systems capable of carrying out the functionality described herein.

Referring to FIG. 20, an example computer system 1904 includes one or more processors, such as processor 2004. Processor 2004 is connected to a communication bus 2002. Various software embodiments are described in terms of this example computer system 1904. After reading this description, it will become apparent to a person skilled in the relevant art how to implement the invention using other computer systems and/or computer architectures.

Computer system 1904 also includes a main memory 2006, preferably random access memory (RAM), and can also include a secondary memory 2008. Secondary memory 2008 can include, for example, a hard disk drive 2010 and/or a removable storage drive 2012, representing a floppy disk drive, a magnetic tape drive, an optical disk drive, etc. Removable storage drive 2012 reads from and/or writes to a removable storage unit 2014 in a well known manner. Removable storage unit 2014, represents a floppy disk, magnetic tape, optical disk, etc. which is read by and written to by removable storage drive 2012. Removable storage unit 2014 includes a computer usable storage medium having stored therein computer software and/or data.

In alternative embodiments, secondary memory 2008 can include other similar means for allowing computer programs or other instructions to be loaded into computer system 1904. Such means can include, for example, a removable storage unit 2022 and an interface 2020. Examples of such can include a program cartridge and cartridge interface (such as that found in video game devices), a removable memory chip (such as an EPROM, or PROM) and associated socket, and other removable storage units 2022 and interfaces 2020 which allow software and data to be transferred from the removable storage unit 2022 to computer system 1904.

Computer system 1904 can also include a communications interface 2024. Communications interface 2024 allows software and data to be transferred between computer system 1904 and external devices. Examples of communications interface 2024 include, but are not limited to a modem, a network interface (such as an Ethernet card), a communications port, a PCMCIA slot and

card, etc. Software and data transferred via communications interface 2024 are in the form of signals which can be electronic, electromagnetic, optical or other signals capable of being received by communications interface 2024. These signals 2026 are provided to communications interface via a channel 2028. This channel 2028 carries signals 2026 and can be implemented using wire or cable, fiber optics, a phone line, a cellular phone link, an RF link and other communications channels.

In this document, the terms "computer program medium" and "computer usable medium" are used to generally refer to media such as removable storage device 2012, a hard disk installed in hard disk drive 2010, and signals 2026. These computer program products are means for providing software to computer system 1904.

Computer programs (also called computer control logic) are stored in main memory and/or secondary memory 2008. Computer programs can also be received via communications interface 2024. Such computer programs, when executed, enable the computer system 1904 to perform the features of the present invention as discussed herein. In particular, the computer programs, when executed, enable the processor 2004 to perform the features of the present invention. Accordingly, such computer programs represent controllers of the computer system 1904.

In an embodiment where the invention is implemented using software, the software can be stored in a computer program product and loaded into computer system 1904 using removable storage drive 2012, hard drive 2010 or communications interface 2024. The control logic (software), when executed by the processor 2004, causes the processor 2004 to perform the functions of the invention as described herein.

In another embodiment, the present invention is implemented primarily in hardware using, for example, hardware components such as application specific integrated circuits (ASICs). Implementation of the hardware state machine so as

to perform the functions described herein will be apparent to persons skilled in the relevant art(s).

In yet another embodiment, the invention is implemented using a combination of both hardware and software.

Conclusion

The previous description of the preferred embodiments is provided to enable any person skilled in the art to make or use the present invention. While the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention. For example, similarity and diversity (dissimilarity) are only two of the many criteria (i.e., fitness functions) that can be used in the present invention. Additional criteria include, but are not limited to, desired properties or property distributions, 2D and 3D QSAR predictions, and receptor complementarity.

The present invention has been described above with the aid of functional building blocks illustrating the performance of specified functions and relationships thereof. The boundaries of these functional building blocks have been arbitrarily defined herein for the convenience of the description. A number of combinatorial synthesis strategies employing various functional building blocks and synthesis strategies for assembling libraries therefrom are described in the art. Alternate boundaries can be defined so long as the specified functions and relationships thereof are appropriately performed. Any such alternate boundaries are thus within the scope and spirit of the claimed invention. One skilled in the art will recognize that these functional building blocks can be implemented by discrete components, application specific integrated circuits, processors executing appropriate software and the like or any combination thereof. Thus, the breadth and scope of the present invention should not be limited by any of the above-

described exemplary embodiments, but should be defined only in accordance with the following claims and their equivalents.